Predicting Bitcoin Price and Trend Forecasting Utilizing a Hybrid Approach of Financial and Textual Data

Rough draft

Eleftherios Kanavakis
ekanavakis@ethz.ch
ETH Zurich
Switzerland

The aim of this report is to give a detailed overview of the data, methodology, results and final steps of this project.

1 Introduction

This paper introduces a unique, integrative approach to Bitcoin (BTC) price and trend prediction by exploiting both financial and textual data. Bitcoin, as the leading cryptocurrency, plays a critical role in the financial market, making its price prediction an essential task. Traditional methods for predicting Bitcoin's price heavily rely on the analysis of historical financial data and trading indicators, such as moving averages. However, in an era characterized by information overflow and the widespread influence of social media on markets, it is reasonable to assume that public sentiment, as expressed through various social media platforms, has a significant impact on Bitcoin price fluctuations.

Consequently, this study proposes a hybrid prediction model that merges traditional financial indicators with a sentiment analysis of Bitcoin-related tweets. For the financial data, we leverage historical Bitcoin prices and trading indicators, which provides a quantitative perspective of market behavior. In contrast, for textual data, we utilize sentiment analysis on tweets related to Bitcoin to incorporate the qualitative aspect of public sentiment. This approach captures the essential factors contributing to the price changes in Bitcoin and improves the forecasting capability of the prediction model. By combining these two diverse but complementary sources of data, we aim to enhance the accuracy of BTC price and trend forecasting and to provide a more comprehensive understanding of the driving forces behind Bitcoin's market movements.

2 Dataset

The financial dataset that we are utilizing for this project originates from Bitfinex[2], one of the industry's leading cryptocurrency exchange platforms. The dataset covers a period from May 2018 to June 2023, equating to 1859 days or 44635 hourly candles. This time range is particularly interesting as it includes Bitcoin's recent bear market, the first to be significantly influenced by global macroeconomic factors since Bitcoin's inception, thus providing a rich context for our analysis. The dataset comprises the following key features:

- Open price: The opening price of Bitcoin at the start of each hourly candle.
- Close price: The closing price of Bitcoin at the end of each hourly candle.
- High: The maximum price Bitcoin reached within the duration of each hourly period.
- Low: The minimum price Bitcoin dipped to within each hourly interval.
- Date: The precise timestamp corresponding to each hourly candle.
- Volume BTC: The total volume of Bitcoin traded during each hourly period.
- Volume USD: The total volume of USD traded against the Bitcoin pair during each hourly period.

With the lowest recorded Bitcoin price at $3,215.2 and the highest soaring to $68,958, the dataset effectively illustrates the inherent volatility and unpredictability of the Bitcoin market. It is our belief that this comprehensive dataset, enhanced by the inclusion of a unique bear market scenario influenced by macroeconomic factors, can provide invaluable insights to develop a reliable model for Bitcoin price and trend prediction.

To complement our understanding, we've visually plotted1 the Bitcoin data over time, distinctly segregating the dataset into training and testing subsets, aiding in both our historical comprehension and future price and trend predictions.

The textual data supplementing our financial dataset is derived from Twitter[1], an influential social media platform allowing users to post succinct messages or 'tweets'. Limited to 280 characters, these tweets encapsulate personal perspectives, news updates, opinions, and discussions on a broad spectrum of topics, including Bitcoin and other cryptocurrencies.

The Twitter dataset extends from February 5, 2021, to March 5, 2023. This duration enables us to glean a comprehensive overview of public sentiment towards Bitcoin over

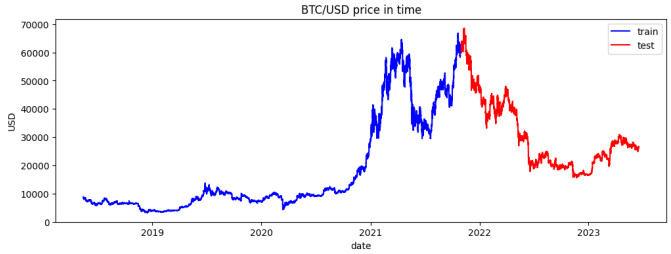

Figure 1. Bitcoin historical data in time

recent years. Encompassing a considerable count of 4,689,288 tweets, this dataset portrays a diverse range of opinions, dialogues, news, and sentiment concerning Bitcoin.

The data features fall into two broad categories:

User-Related Information: This includes aspects such as user_name, user_location, user_description, user_created, user_followers, user_friends, user_favourites, and user_verified, providing a broad context of the user profiles interacting with Bitcoin-related content.

Tweet-Related Information: This pertains to details like date, text, hashtags, source, and is_retweet, capturing the essence, timing, reach, and sentiment of the tweets.

Through the analysis of this rich dataset, we plan to assess the impact of public sentiment on Bitcoin's price and trend dynamics, supplementing the quantitative financial data with these qualitative insights.

3 Methodology

In this section, we adopt distinct modeling approaches for our two types of data: financial and textual. For the financial data, we utilize the historical price and volume trends of Bitcoin, while for the textual data, we perform sentiment analysis on relevant tweets. Finally, we merge these models to construct a comprehensive predictive model that capitalizes on both the quantitative insights from the financial data and the qualitative sentiment indicators from the Twitter data. This combined approach aims to offer a more holistic and accurate prediction of Bitcoin's price and market trends.

3.1 Financial Data Modelling

3.1.1 Tasks. In the context of the financial data, we will focus on two main tasks: Bitcoin (BTC) trend prediction and Bitcoin price prediction.

- Bitcoin Trend Prediction: This task is essentially a binary classification problem. Our goal here is to predict the close trend for timestep t+1, based on timestep t. In practical terms, if the closing price at timestep t is less than the closing price at timestep t+1, we consider it as an uptrend. Conversely, if the closing price at timestep t is greater than or equal to the closing price at timestep t+1, we interpret this as a downtrend.
- Bitcoin Price Prediction: The second task involves predicting the Bitcoin price, which is a regression problem. Here, the objective is to predict the closing price at timestep t+1, given the information available at timestep t.

To evaluate the performance of our models, we will utilize metrics inspired by related work in the field. For the task of trend prediction, we will apply the F1 score, a balanced measure of precision and recall. For price prediction, we will employ the Square Root of the Mean Squared Error (RMSE), a standard measure for regression tasks that assesses the average magnitude of prediction errors.

Both of these tasks contribute to our larger goal of developing a comprehensive predictive model for Bitcoin's price and market trends, taking into account not only these financial indicators but also the sentiment data extracted from Twitter.

3.1.2 Data Processing. In our data processing methodology and experiments, we employ three different strategies to build, compare, and improve our predictive models for Bitcoin price and trend.

Baseline Model: For our baseline, we utilize the "vanilla" or raw data as it is. This involves feeding the financial data, without any additional transformations or feature engineering, into our models to establish a benchmark for their performance.

Time-Specific Feature Addition: Building on our baseline, our second approach involves augmenting our vanilla data with time-specific features, such as the day of the week, month, and year. The rationale behind this strategy is grounded in the recurrent patterns often observed in financial markets, where certain behaviors tend to repeat on specific days, weeks, or months. Incorporating this temporal information could potentially enhance the learning capability of our models.

Combined Features Model: Our third and final strategy involves fusing the original features (technical indicators) with specific trading indicators. For the extraction and computation of these trading indicators, we leverage the Stockstats library. The technical indicators used include, but are not limited to, the following:

- Relative Strength Index (RSI): A momentum oscillator that measures the speed and change of price movements.
- Moving Average Convergence Divergence (MACD): A trend-following momentum indicator that shows the relationship between two moving averages of a security's price.
- Bollinger Bands: A volatility indicator that consists of a simple moving average (SMA) alongside upper and lower bands based on standard deviations.

- Average True Range (ATR): A technical analysis volatility indicator originally developed by J. Welles Wilder, Jr.
- Commodity Channel Index (CCI): A momentum-based oscillator used to help determine when an investment vehicle is reaching a condition of being overbought or oversold.

Through these three strategies, we aim to uncover the most effective approach for predicting Bitcoin's price and market trends, providing a comprehensive comparison and understanding of the different methods and their performances.

3.1.3 Modelling. To build the most robust predictive models, we'll be utilizing three varied architectures: XGBoost, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) networks. Before modeling, we ensure all input features are normalized to a similar scale, enhancing model training efficiency.

XGBoost, a gradient-boosting framework known for its effectiveness, will be applied as both a classifier and a regressor for trend and price prediction tasks respectively. Meanwhile, MLP, a type of artificial neural network, and LSTM, a form of recurrent neural network, will be utilized in both tasks. Configured as either classifiers or regressors as needed, they will work on predicting Bitcoin's market trend and price. These models collectively provide a comprehensive framework for tackling our prediction tasks.

3.2 Sentiment Analysis

3.2.1 Data Processing. For the textual data sourced from Twitter, we apply specific filtering criteria to streamline the dataset and focus on the most relevant information for our task. Our initial step involves retaining only those tweets that contain Bitcoin-related hashtags such as 'btc', 'bitcoin', or 'Bitcoin'. This not only enables us to hone in on the most pertinent discussions and sentiments about Bitcoin but also helps in managing the vast volume of data, given our processing resources.

Additionally, we restrict our dataset to include only tweets that originate from mobile devices. This decision is motivated by two key reasons. Firstly, mobile-originated tweets are often less prone to being generated by automated bots, thereby reducing the noise in our data. Secondly, this filter, much like the hashtag criteria, contributes to our data pruning efforts. Through these measures, we ensure our textual data is both manageable and highly relevant to Bitcoin trend and price prediction.

Having pared down our initial dataset to a manageable volume of relevant tweets, we now turn our attention to the text processing phase. This stage is essential in preparing our data for the subsequent modeling. Here we refine the content of the tweets, a process integral to the efficacy of our sentiment analysis model.

Initially, we discarded the hashtag symbol (#) to focus solely on the textual substance of the tags. Following that, we employed regular expressions to remove hyperlinks from the tweets, as these do not contribute to the overall sentiment conveyed by the text. Finally, we extricated all mentions from the tweets, which are typically usernames prefixed with '@'. This allows our model to focus on the key text content without being swayed by user-specific data. These meticulous cleaning steps resulted in a curated dataset of 891,229 tweets, primed for effective sentiment analysis.

3.2.2 Modelling. For sentiment analysis, we utilize the 'distilbert-base-uncased-finetuned-sst-2-english' model, which is part of the Hugging Face's transformers library. This model represents a distilled and optimized version of the robust BERT model, successfully maintaining 95% of BERT's performance capabilities while being 60% smaller and faster. This makes it a high-performing yet efficient tool for our requirements.

What's more, the model has been specifically fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset, a resource frequently used for sentiment analysis tasks. Through this fine-tuning, the model has honed its ability to categorize sentences into positive or negative sentiment, a skill directly applicable to our goal of assessing sentiment within Bitcoin-related tweets.

In our process, we feed the tweet text data into this fine-tuned model, which subsequently outputs a sentiment score for each tweet. This method provides us with a potent mechanism for gleaning sentiment insights from a substantial corpus of tweet data, thereby enriching our understanding of public sentiment around Bitcoin during the study period.

3.3 Combination of methods

To integrate the sentiment analysis results with our forecasting models, we devised a process to coalesce these two diverse streams of data. Our strategy involves calculating the hourly average sentiment score for each Bitcoin-related tweet, resulting in an overall sentiment score per hour. This approach effectively quantifies the average public sentiment towards Bitcoin for every hour in our study period.

The derived hourly sentiment scores are then added as an additional feature column to the financial dataset. This incorporation of sentiment information allows our forecasting models to utilize not only the financial indicators but also the prevailing public sentiment towards Bitcoin during each hour, thereby providing a more comprehensive understanding for the forecasting task.

In scenarios where sentiment scores are absent or where a robust prediction could not be made, we assign a neutral value. This decision ensures continuity in our dataset, allowing the models to function without interruption due to missing or unreliable sentiment data.

4 Results

4.1 Financial Data

In this section, we present the results obtained from our analysis of financial data, on the tasks of trend prediction and price forecasting for Bitcoin. These results are critical as they shed light on the effectiveness of both our data processing and chosen models in predicting Bitcoin's future behavior, based on its historical financial data.

Data	Model	Trend(F1)	Price(RMSE)
Vanilla	XGBoost	0.49	821.51
Vanilla	MLP	0.49	238.98
Vanilla	LSTM	0.54	255.75
Vanilla + Time	XGBoost	0.49	828.77
Vanilla + Time	MLP	0.49	238.98
Vanilla + Time	LSTM	0.54	255.75
Vanilla + Trading	XGBoost	0.50	818.52
Vanilla + Trading	MLP	0.49	238.98
Vanilla + Trading	LSTM	0.57	255.76

Table 1. Results for trend prediction and price prediction tasks using different data experiments across various models.

Building on these observations1, it becomes clear that our efforts in diversifying the feature sets and exploring multiple models have generated several valuable insights. Firstly, despite the theoretical appeal of trading indicators and time-specific features, they failed to significantly boost our models' predictive performances, likely due to the inherent complexity of the test period characterized by an intense bear market.

When it comes to model comparison, XGBoost and MLP presented comparable performances for the trend prediction, whereas LSTM slightly pulled ahead, indicating its superior capability in managing complex temporal dependencies. In the price prediction task, MLP and LSTM managed to achieve lower RMSE values, suggesting their potential edge in handling continuous prediction tasks. However, all models were equally challenged by the unpredictable nature of the test period, a testament to the fundamental difficulties involved in financial forecasting.

To complement our understanding, we've visually plotted2 the Bitcoin data over time, and the results from our forecasting model.

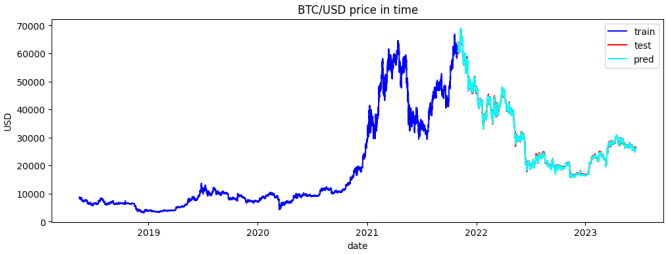

Figure 2. Bitcoin price forecasting results

4.2 Financial and Textual Data

Our project is currently in an interim stage where the integration of financial and sentiment analysis data into our machine learning models is still ongoing. We expect to conclude these comprehensive experiments soon, with the results enriching our understanding of Bitcoin price and trend forecasting. The finalized findings will be presented in an upcoming version of this paper.

References

[1] [n.d.]. Bitcoin Tweets. https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweet

[2] [n.d.]. BTC/USD Historical Data. https://www.cryptodatadownload.com/data/bitfinex/